Title

Universal adversarial perturbations for speech recognition systems

Authors

Neekhara, P
Hussain, S
Pandey, P
et al.

Publication Date

2019

DOI

10.21437/Interspeech.2019-1353

Peer reviewed

Universal Adversarial Perturbations for Speech Recognition Systems

*Paarth Neekhara[1], *Shehzeen Hussain[2], Prakhar Pandey[1], Shlomo Dubnov[3], Julian McAuley[1], Farinaz Koushanfar[2]

[1]UC San Diego Department of Computer Science
[2]UC San Diego Department of Electrical and Computer Engineering
[3]UC San Diego Department of Music
* Equal contribution

pneekhar@ucsd.edu, ssh028@ucsd.edu

Abstract

In this work, we demonstrate the existence of universal adversarial audio perturbations that cause mis-transcription of audio signals by automatic speech recognition (ASR) systems. We propose an algorithm to find a single quasi-imperceptible perturbation, which when added to any arbitrary speech signal, will most likely fool the victim speech recognition model. Our experiments demonstrate the application of our proposed technique by crafting audio-agnostic universal perturbations for the state-of-the-art ASR system – Mozilla DeepSpeech. Additionally, we show that such perturbations generalize to a significant extent across models that are not available during training, by performing a transferability test on a WaveNet based ASR system.

Index Terms: speech recognition, adversarial examples, speech processing, computer security

1. Introduction

Machine learning agents serve as the backbone of several speech recognition systems, widely used in personal assistants of smartphones and home electronic devices (e.g. Apple Siri, Google Assistant). Traditionally, Hidden Markov Models (HMMs) [1, 2, 3, 4, 5, 6] were used to model sequential data but with the advent of deep learning, state-of-the-art speech recognition systems are based on Deep Neural Networks (DNNs) [7, 8, 9, 10].

However, several studies have demonstrated that DNNs are vulnerable to adversarial examples [11, 12, 13, 14, 15]. An adversarial example is a sample from the classifier's input domain which has been perturbed in a way that is intended to fool a victim machine learning (ML) model. While the perturbation is usually imperceptible, such an adversarial input can mislead neural network models deployed in real-world settings causing it to output an incorrect class label with higher confidence.

A vast amount of past research in adversarial machine learning has shown such attacks to be successful in the image domain [16, 15, 17, 18, 19, 20]. However, few works have addressed attack scenarios involving other modalities such as audio. This limits our understanding of system vulnerabilities of many commercial speech recognition models employing DNNs, such as Amazon Alexa, Google Assistant, and home electronic devices like Amazon Echo and Google Home. Recent studies that have explored attacks on automatic speech recognition (ASR) systems [21, 22, 23, 24], have demonstrated that adversarial examples exist in the audio domain. The authors of [22] proposed targeted attacks where an adversary designs a perturbation that can cause the original audio signal to be transcribed to any phrase desired by the adversary. However, calculating such perturbations requires the adversary to solve an optimization problem for each data-point they wish to mis-transcribe. This makes the attack in-applicable in real-time since the adversary would need to re-solve the data-dependent optimization problem from scratch for every new data-point.

Universal Adversarial Perturbations [25] have demonstrated that there exist universal *image-agnostic* perturbations which when added to any image will cause the image to be misclassified by a victim network with high probability. The existence of such perturbations poses a threat to machine learning models in real world settings since the adversary may simply add the same pre-computed universal perturbation to a new image and cause mis-classification.

Contributions: In this work, we seek to answer the question "Do universal adversarial perturbations exist for neural networks in audio domain?" We demonstrate the existence of universal audio-agnostic perturbations that can fool DNN based ASR systems.[1] We propose an algorithm to design such universal perturbations against a victim ASR model in the *white-box setting*, where the adversary has access to the victim's model architecture and parameters. We validate the feasibility of our algorithm, by crafting such perturbations for Mozilla's open source implementation of the state-of-the-art speech recognition system DeepSpeech [10]. Additionally, we discover that the generated universal perturbation is transferable to a significant extent across different model architectures. Particularly, we demonstrate that a universal perturbation trained on DeepSpeech can cause significant transcription error on a WaveNet [9] based ASR model.

2. Related Work

Adversarial Attacks in the Audio Domain: Adversarial attacks on ASR systems have primarily focused on *targeted attacks* to embed carefully crafted perturbations into speech signals, such that the victim model transcribes the input audio into a specific malicious phrase, as desired by the adversary [21, 22, 26, 23, 27]. Prior works [23, 27] demonstrate successful attack algorithms targeting traditional speech recognition models based on HMMs and GMMs, that operate on Mel Frequency Cepstral Coefficient (MFCC) representation of audio. For example, in Hidden Voice Commands [23], the attacker uses inverse feature extraction to generate obfuscated audio that can be played over-the-air to attack ASR systems. However, obfuscated samples sound like random noise rather than normal human perceptible speech and therefore come at the cost of being

[1]Sound Examples: universal-audio-perturbation.herokuapp.com

fairly perceptible to human listeners. Additionally, these attack frameworks are not end-to-end, which render them impractical for studying the vulnerabilities of modern ASR systems – that are entirely DNN based.

In more recent work [22], Carlini *et al.* propose an end-to-end white-box attack technique to craft adversarial examples, which transcribe to a target phrase. Similar to the work in images, they propose a gradient-based optimization method that replaces the cross-entropy loss function used for classification, with a Connectionist Temporal Classification (CTC) loss [28] which is optimized for time-sequences. The CTC-loss between the target phrase and the network's output is backpropagated through the victim neural network and the MFCC computation, to update the additive adversarial perturbation. The adversarial samples generated by this work are quasi-perceptible, motivating a separate work [29] to minimize the perceptibility of the adversarial perturbations using psychoacoustic hiding.

Designing adversarial perturbations using all the above mentioned approaches requires the adversary to solve a data dependent optimization problem for each input audio signal the adversary wishes to mis-transcribe, making them ineffective in a real-time attack scenario. In other words, targeted attacks must be customized for each segment of audio, a process that cannot yet be done in real-time. The existence of universal adversarial perturbations (described below) can pose a more serious threat to ASR systems in real-world settings since the adversary may simply add the same pre-computed universal adversarial perturbation to any input audio and fool the DNN based ASR system.

Universal Adversarial Perturbations: The authors of [25] craft a single universal perturbation vector which can fool a victim neural network to predict a false classification output on the majority of validation instances. Let $\hat{k}(x)$ be the classification output for an input x that belongs to a distribution μ. The goal is to find a perturbation v such that: $\hat{k}(x+v) \neq \hat{k}(x)$ for "most" $x \in \mu$. This is formulated as an optimization problem with constraints to ensure that the universal perturbation is within a specified p-norm and is also able to fool the desired number of instances in the training set. The proposed algorithm iteratively goes over the training dataset to build a universal perturbation vector that pushes each data point to its decision boundary. The authors demonstrate that it is possible to find a quasi-imperceptible universal perturbation that pushes most data points outside the correct classification region of a victim model. More interestingly, the work demonstrates that the universal perturbations are transferable across models with different architectures. The perturbation produced using one network such as VGG-16 can also be used to fool another network e.g. GoogLeNet showing that their method is doubly universal. *Universal adversarial perturbations* for images focuses on the goal of mis-classification and cannot directly be applied to the more challenging goal of mis-transcription by Speech Recognition System. In our work we address this challenge and solve an alternate optimization problem to adapt the method for designing universal adversarial perturbations for ASR systems.

3. Methodology

3.1. Threat Model

We aim to find a universal audio perturbation, which when added to any speech waveform, will cause an error in transcription by a speech recognition model with high probability. For the success of the attack, the error in the transcription should

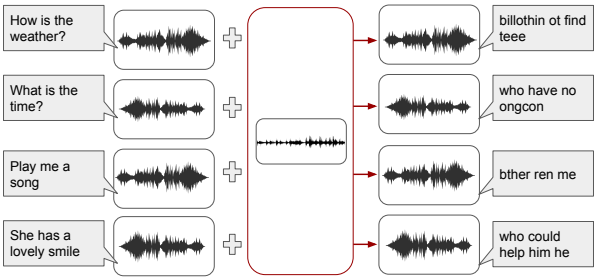

Figure 1: *Threat Model: We aim to find a single perturbation which when added to any arbitrary audio signal, will most likely cause an error in transcription by a victim Speech Recognition System*

be high enough so that the transcription of the perturbed signal (adversarial transcription) is incomprehensible and the original transcription cannot be deduced from the adversarial transcription. As discussed in [22], the transcription *"test sentence"* misspelled as *"test sentense"* does little to help the adversary. To make the adversary's goal challenging, we report success only when the Character Error Rate (CER) or the normalized Levenshtein distance *(Edit Distance)* [30] between the original and adversarial transcription is greater than a particular threshold. Formally, we define our threat model as follows:

Let μ denote a distribution of waveforms and C be the victim speech recognition model that transcribes a waveform x to $C(x)$. The goal of our work is to find perturbations v such that:

$$CER(C(x), C(x+v)) > t \text{ for "most" } x \in \mu$$

Here, $CER(x, y)$ is the edit distance between the strings x and y normalized [30] by the $length$ of x i.e

$$CER(x, y) = \frac{EditDistance(x, y)}{length(x)}$$

The threshold t is chosen as 0.5 for our experiments i.e., we report success only when the original transcription has been *edited* by at least 50% of its length using character *removal, insertion, or substitution* operations.

The universal perturbation signal v is chosen to be of a fixed length and is cropped or zero-padded at the end to make it equal to length of the signal x.

3.2. Distortion Metric

To quantify the distortion introduced by some adversarial perturbation v, an l_∞ metric is commonly used in the space of images. Following the same convention, in the audio domain [13], the loudness of the perturbation can be quantified using the dB scale, where $dB(v) = \max_i(20.\log_{10}(v_i))$. We calculate $dB_x(v)$ to quantify the relative loudness of the universal perturbation v with respect to an original waveform x where:

$$dB_x(v) = dB(v) - dB(x)$$

Since the perturbation introduced is quieter than the original signal, $dB_x(v)$ is a negative value, where smaller values indicate quieter distortions. In our results, we report the average relative loudness: $dB_x(v)$ across the whole test set to quantify the distortion introduced by our universal perturbation.

3.3. Problem Formulation and Algorithm

Our goal to find a quasi-imperceptible universal perturbation vector v such that it mis-transcribes *most* data points sampled from a distribution μ. Mathematically, we want to find a perturbation vector v that satisfies:

1. $\|v\|_\infty < \epsilon$
2. $\underset{x\sim\mu}{P}(CER(C(x), C(x+v) > t)) \geq \delta$.

Here ϵ is the maximum allowed l_∞ norm of the perturbation, δ is the desired attack success rate and t is the threshold CER chosen to define our success criteria.

To solve the above problem, we adapt the universal adversarial perturbation algorithm proposed by [25] to find universal adversarial perturbations for the goal of *mis-transciption* of speech waveforms instead of *mis-classification* of data (images). Let $X = x_1, x_2, \ldots, x_m$ be a set of speech signals sampled from the distribution μ. Our Algorithm (1) goes over the data-points in X iteratively and gradually builds the perturbation vector v. At each iteration i, we seek a minimum perturbation Δv_i, that causes an error in the transcription of the current perturbed data point $x_i + v$. We then add this additional perturbation Δv_i to the current universal perturbation v and clip the new perturbation v, if necessary, to satisfy the constraint $\|v\|_\infty < \epsilon$.

Algorithm 1 Universal Adversarial Perturbations for Speech Recognition Systems

input: Training Data Points X, Validation Data Points X_v Victim Model C, allowed distortion level ϵ, desired success rate δ
output: Universal Adversarial Perturbation vector v
Initialize $v \leftarrow 0$
while $SuccessRate(X_v) < \delta$ **do**
  for each data point $x_i \in X$ **do**
    if $CER(C(x_i + v + r), C(x_i)) < t$ **then**
      Compute min perturbation that mis-transcribes $x_i + v$:
      $\Delta v_i \leftarrow \arg\min_r \|r\|_2$ s.t.:
      $CER(C(x_i + v + r), C(x_i)) > t$
      Update and clip universal perturbation v:
      $v = Clip_{v,\epsilon}(v + \Delta v_i)$

At each iteration we need to solve the following optimization problem, that seeks a minimum (under *l2* norm) additional perturbation Δv_i, to mis-transcribe the current perturbed audio signal $x_i + v$:

$$\Delta v_i \leftarrow \arg\min_r \|r\|_2 \text{ s.t. } CER(C(x_i + v + r), C(x_i)) > t \tag{1}$$

It is non-trivial to solve the above optimization in its current form. In [25], the authors try to solve a similar optimization problem for the goal of *mis-classification* of data points. They approximate its solution using DeepFool [31] which finds a minimum perturbation vector that pushes a data point to its decision boundary. Since we are tackling a more challenging goal of *mis-transcription* of signals where we have decision boundaries for each audio frame across the time axis, the same idea cannot be directly applied. Therefore, we approximate the solution to the optimization problem given by Equation 1 by solving a more tractable optimization problem:

$$\begin{aligned} &\text{Minimize } J(r) \text{ where} \\ &J(r) = c\|r\|^2 + L(x_i + v + r, C(x_i)) \\ &\text{s.t. } \|v + r\|_\infty < \epsilon \\ &\text{where } L(x, y) = -CTCLoss(f(x), y) \end{aligned} \tag{2}$$

In other words, to mis-transcribe the signal, we aim to maximize the CTC-Loss between the predicted probability distributions of the perturbed signal $f(x_i + v + r)$ and the original transcription $C(x_i)$ while having a regularization penalty on the $l2$ norm of r. Since this a non-convex optimization problem, we approximate its solution using iterative gradient sign method [32]:

$$\begin{aligned} r_0 &= \overrightarrow{0} \\ r_{N+1} &= Clip_{r+v,\epsilon}\{r_N - \alpha sign(\Delta_{r_N} J(r_N)\} \end{aligned} \tag{3}$$

Note that the error J is back-propagated through the entire neural network and the MFCC computation to the perturbation vector r. We iterate until we reach the desired CER threshold t for a particular data point x_i. The regularization constant c is chosen through hyper-parameter search on a validation set to find the maximum success rate for a given magnitude of allowed perturbation.

4. Experimental Details

We demonstrate the application of our proposed attack algorithm on the pre-trained *Mozilla DeepSpeech* model [33, 10]. We train our algorithm on the Mozilla Common Voice Dataset [10] which contains 582 hours of audio across 400,000 recordings in English. We train on a randomly selected set *X* containing 5,000 audio files from the training set and evaluate our model on both the training set X and the entire unseen validation set of the Mozilla Common Voice Dataset. We analyze the effect of the size of the set X below. The length of our universal adversarial perturbation is fixed to 150,000 samples which corresponds to around 9 seconds of audio at 16 KHz. The universal adversarial perturbations are trained using our proposed algorithm 1 with a learning rate $\alpha = 5$ and the regularization parameter c set to 0.5.

Evaluation: We utilize two metrics: *i) Mean CER* - Character Error Rate averaged over the entire test set and *ii) Success Rate* to evaluate our universal adversarial perturbations. We report success on a particular waveform, if the *CER* between the original and adversarial transcription (Section 3.1) is greater than 0.5. The amount of perturbation is quantified using mean relative distortion $dB_x(v)$ over the test set (Refer to Section 3.2).

5. Results

Table 1 shows the results of our algorithm for different allowed magnitude of universal adversarial perturbation on both the training set X and the unseen Test Set. Both the success rate and the Mean Character Error Rate (CER) increase with increase in the maximum allowed perturbation. We achieve a success rate of 89.06 % on the validation set, with the mean distortion metric $dB_x(v) \approx -32dB$. To interpret the results in context, $-32dB$ is roughly the difference between ambient noise in a quiet room and a person talking [34, 22] . We encourage the reader to listen to our adversarial samples and their corresponding transcriptions on our web page (link in the footnote of the first page)

Table 1: *Results of our algorithm for different allowed magnitude of universal adversarial perturbation*

	Training Set (X)			Test Set		
$\|\mathbf{v}\|_\infty$	Mean $\mathbf{dB_x(v)}$	Success Rate (%)	Mean CER	Mean $\mathbf{dB_x(v)}$	Success Rate (%)	Mean CER
100	-42.03	57.46	0.63	-41.86	56.13	0.64
150	-38.51	72.78	0.81	-38.34	72.49	0.82
200	-36.01	83.27	0.92	-35.84	80.47	0.95
300	-32.49	89.52	1.10	-32.32	89.06	1.11
400	-30.18	90.60	1.06	-29.82	88.24	1.07

Figure 2 shows the success rate and mean edit distance compared to the size of the training set X for maximum allowed perturbation $\|v\|_\infty = 200$ (Mean $dB_x(v) = -36.01$). We observe that it is possible to train our proposed algorithm on very few examples and achieve reasonable success rates on unseen data. For example, training on just 1000 examples can achieve a success rate of 80.47 % on the test set.

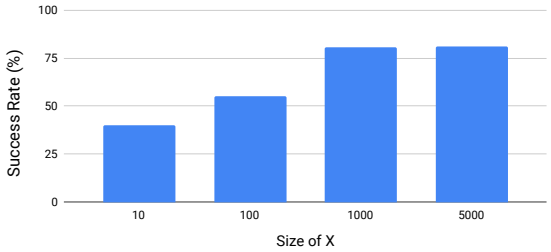

Figure 2: *Attack Success Rate on the test set vs. the number of audio files in the training set X*

5.1. Effectiveness of universal perturbations

In order to assess the vulnerability of the victim Speech Recognition System to our attack algorithm, we compare our universal perturbation with random (uniform) perturbation having the same magnitude of distortion (same $\|v\|_\infty$) as our universal adversarial perturbation. Figure 3 shows the plot of success rate vs. the magnitude of the perturbation for each of these perturbations. It can be seen that universal adversarial perturbations are able to achieve high success rate with very low magnitude of distortion as compared to a random noise perturbation. For example, for allowed perturbation $\|v\|_\infty = 100$ our universal perturbation achieves a success rate of 65% which is substantially higher than the success rate of random noise. This implies that for the same magnitude of distortion, distorting an audio waveform in a random direction is significantly less likely to cause mis-transcription as compared to distorting the waveform in the direction of universal perturbation. Our results support the hypothesis discussed in [25], demonstrating that universal adversarial perturbations exploit geometric correlations in the decision boundaries of the victim model.

5.2. Cross-model Transferability

We perform a study on the transferability of adversarial samples to deceive ML models that have not been used for training the universal adversarial perturbation, i.e., their parameters and network structures are not revealed to the attacker. We train universal adversarial perturbations for Mozilla DeepSpeech and evaluate the extent to which they are valid for a different ASR

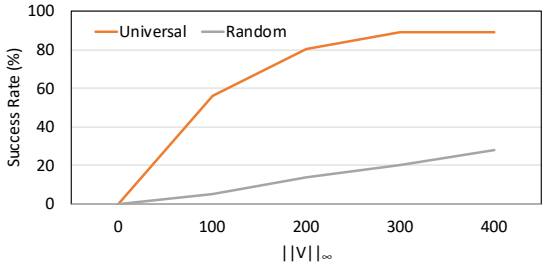

Figure 3: *Success Rate vs $\|v\|_\infty$ of universal and random perturbations.*

Table 2: *Results of the same universal adversarial perturbation on two victim models: Wavenet and Mozilla DeepSpeech. The universal perturbation was trained on the DeepSpeech model.*

		Wavenet		Mozilla DeepSpeech	
$\|\mathbf{v}\|_\infty$	Mean dBx(v)	Success Rate (%)	Mean CER	Success Rate (%)	Mean CER
150	-38.34	**26.97**	**0.37**	72.49	0.82
200	-35.84	**31.18**	**0.40**	80.47	0.95
300	-32.32	**42.05**	**0.47**	89.06	1.11
400	-29.82	**63.28**	**0.60**	88.24	1.07

architecture based on WaveNet [9]. For this study, we use a publicly available pre-trained model of WaveNet [35] and evaluate the transcriptions obtained using clean and adversarial audio for the same unseen validation dataset as used in our previous experiments. Our results in Table 2 indicate that our attack is transferable to a significant extent for this particular setting. Specifically, when the mean $dB_x(v) = -29.82$, we are able to achieve a 63.28% success rate while attacking the WaveNet based ASR model. This result demonstrates the practicality of such adversarial perturbations, since they are able to generalize well across data points and architectures.

6. Conclusion

In this work, we demonstrate the existence of audio-agnostic adversarial perturbations for speech recognition systems. We demonstrate that our audio-agnostic adversarial perturbation generalizes well across unseen data points and to some extent across unseen networks. Our proposed end-to-end approach can be used to further understand the vulnerabilities and blind spots of deep neural network based ASR system, and provide insights for building more robust neural networks.

7. References

[1] L. E. Baum and J. A. Eagon, "An inequality with applications to statistical estimation for probabilistic functions of markov processes and to a model for ecology," *Bull. Amer. Math. Soc.*, vol. 73, no. 3, pp. 360–363, 1967.

[2] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains," *The annals of mathematical statistics*, vol. 41, no. 1, pp. 164–171, 1970.

[3] A. Acero, l. Deng, T. Kristjansson, and J. Zhang, "Hmm adaptation using vector taylor series for noisy speech recognition," 01 2000, pp. 869–872.

[4] S. Ahadi and P. C. Woodland, "Combined bayesian and predictive techniques for rapid speaker adaptation of continuous density

hidden markov models," *Computer speech & language*, vol. 11, no. 3, pp. 187–206, 1997.

[5] L. Bahl, P. Brown, P. de Souza, and R. Mercer, "Maximum mutual information estimation of hidden markov model parameters for speech recognition," in *ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 11. IEEE, 1986, pp. 49–52.

[6] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[7] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*, 2016, pp. 173–182.

[8] A. van den Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. van den Driessche, E. Lockhart, L. Cobo, F. Stimberg, N. Casagrande, D. Grewe, S. Noury, S. Dieleman, E. Elsen, N. Kalchbrenner, H. Zen, A. Graves, H. King, T. Walters, D. Belov, and D. Hassabis, "Parallel WaveNet: Fast high-fidelity speech synthesis," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. Stockholmsmssan, Stockholm Sweden: PMLR, 2018, pp. 3918–3926.

[9] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016.

[10] A. Y. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, "Deep speech: Scaling up end-to-end speech recognition," *CoRR*, vol. abs/1412.5567, 2014.

[11] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *stat*, vol. 1050, p. 20, 2015.

[12] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, Jul. 2018.

[13] N. Carlini and D. A. Wagner, "Towards evaluating the robustness of neural networks," *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57, 2017.

[14] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," *CoRR*, 2016.

[15] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2016, pp. 372–387.

[16] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *ICLR*, vol. abs/1312.6199., 2014.

[17] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*. ACM, 2017, pp. 506–519.

[18] N. Papernot, P. D. McDaniel, and I. J. Goodfellow, "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples," *CoRR*, vol. abs/1605.07277, 2016. [Online]. Available: http://arxiv.org/abs/1605.07277

[19] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," *CoRR*, vol. abs/1712.09665, 2017. [Online]. Available: http://arxiv.org/abs/1712.09665

[20] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations*, 2015.

[21] M. Alzantot, B. Balaji, and M. B. Srivastava, "Did you hear that? adversarial examples against automatic speech recognition," *CoRR*, vol. abs/1801.00554, 2018. [Online]. Available: http://arxiv.org/abs/1801.00554

[22] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *2018 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2018, pp. 1–7.

[23] N. Carlini, P. Mishra, T. Vaidya, Y. Zhang, M. Sherr, C. Shields, D. Wagner, and W. Zhou, "Hidden voice commands," in *25th USENIX Security Symposium (USENIX Security 16)*. Austin, TX: USENIX Association, 2016, pp. 513–530.

[24] H. Yakura and J. Sakuma, "Robust audio adversarial example for a physical attack," *CoRR*, vol. abs/1810.11793, 2018. [Online]. Available: http://arxiv.org/abs/1810.11793

[25] S. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 86–94.

[26] D. Iter, J. Huang, and M. Jermann, "Generating adversarial examples for speech recognition," 2017.

[27] T. Vaidya, Y. Zhang, M. Sherr, and C. Shields, "Cocaine noodles: Exploiting the gap between human and machine speech recognition," in *9th USENIX Workshop on Offensive Technologies (WOOT 15)*. Washington, D.C.: USENIX Association, 2015.

[28] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 369–376.

[29] L. Schönherr, K. Kohls, S. Zeiler, T. Holz, and D. Kolossa, "Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding," *arXiv preprint arXiv:1808.05665*, 2018.

[30] L. Yujian and L. Bo, "A normalized levenshtein distance metric," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1091–1095, Jun. 2007.

[31] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2574–2582.

[32] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *CoRR*, vol. abs/1607.02533, 2016.

[33] "Project deepspeech," https://github.com/mozilla/DeepSpeech.

[34] S. W. Smith, *The Scientist and Engineer's Guide to Digital Signal Processing*. San Diego, CA, USA: California Technical Publishing, 1997.

[35] "Speech to text wavenet," https://github.com/buriburisuri/speech-to-text-wavenet.